Can collaboratories relieve the predicament of the modern scientist?

Larry Rahn, David Leahy, Jim Myers, Carmen Pancerella

The product of scientists is rapidly evolving from archived journal publications and technology spinoffs to systems-level knowledge bases simultaneously useable by scientists, technologists, industry, and stakeholders. The fabric that generates scientific knowledge is moving from independent and individual investigators to Knowledge Grids composed of collaborating multidisciplinary scientists, shared data and tools, unique facilities, and supporting infrastructures. This evolution is challenging each scientist more and more to work in the context of such a 'systems science' environment, forcing them to deal with rapid change, increasing complexity, and scarce resources (data, time, cycles, manpower, and funding).

A number of combustion researchers find themselves over constrained in the current environment and seek new approaches to produce complex collaborative knowledge products. Private collections of data are being annotated and published with technology that facilitates discovery and interoperability (and hopes of attracting more data). New applications are being built to use this data to produce active versions of thermodynamic tables, and these applications are also being offered to others as web services and/or open source. Rare and misunderstood chemical models are to be replaced with model generation services that produce a plethora of well-documented models, each customized as required by its application. Scientists are calling special purpose virtual communities together to develop, evaluate and validate this new knowledge. Elevated standards and new forms of shared knowledge are envisioned to not only improve progress and clarity but are presented to stakeholders to motivate additional funding.

What must we do to ensure that these capabilities collectively emerge as part of a highly functional Knowledge Grid that enables scientists to coherently contribute in a 'systems science' environment as a part of their everyday research? De Roure and Hendler [IEEE Intelligent Systems, vol. 19, no. 1, 2004, pp.65-71] argue that the envisioned Semantic Grid must "…liberate them from the mechanics of e-science so that they can exercise their scientific expertise to generate new science…" Can we take lessons from the perspective of application scientists to set priorities for infrastructure development so this really happens? How can we be successful when everything, even our vision of the future, is evolving? Many diverse challenges have been pointed out, and many are already being competently tackled. From our experience, though, there are perhaps three important aspects that may be worth more discussion in the context of the DOE National Collaboratory Program:

1) Collaborative knowledge products; What are they? What is needed to produce them?
2) Adoptability – how can we encourage new behaviors and put more 'usability' up front?
3) Adaptability - how can we implement tools in a way that enables continuous change?

1) Collaborative knowledge products have the credibility of a community, offer a coherent, high-level perspective that is a context for more detailed information, and can be intelligibly used by many disciplines for multiple purposes. They impact industrial products and processes, decision makers, and, most importantly, other scientists.

a) Knowledge Bases of data, metadata, abstracts, papers, presentations, computer programs, and even interactive tools. They must be discoverable, able to be linked to each other in different ways, and subjected to continuous improvement and community review. What are all the aspects of KB's and how are scientists enabled to construct them?

b) Knowledge Tools contributing to KB construction: Can data management tools enable the evolution of data/metadata from its origin at an experiment or simulation all the way to its role in an archival knowledge base? How can we develop domain ontologies and other enabling KB features from the 'bottom up'? What knowledge extraction and retrieval tools are needed?

c) Supporting middleware: Multiscale sciences develop layers of data that support models that interact in networks or simulations to describe complex phenomena. What is are the technologies that can keep track of such diverse objects and their connectedness? What semantic technologies will facilitate the automated placement and connectedness?

d) Simulation and data analysis codes and unique facilities are part of the Knowledge Grid.

e) What types of advanced visualization and other technologies are needed to help users to absorb knowledge?

f) Independent credible public information sources: Is it possible that fusion with the WWW also makes it possible for a scientific community to offer credible, accessible information to decision makers, including the general public?

2) Adoptability and focus on the end users is critically important if development efforts in collaboratories and knowledge management systems. New knowledge management and collaboration tools generally require new behaviors (and often, modified concepts) for users.

a) This 'newness' is often the biggest barrier to adoption, sometimes requiring shifts in concepts and even requirements perceived by scientists. Examples include electronic security, ownership of data and code, roles and perceived value/recognition of scientists.

b) The tools must offer a 'value proposition' that bears up under a cost-benefit analysis for the scientists. The users must be able to see how the tools will significantly improve their scientific productivity. Enabling collaborative construction of chemical models from distributed data resources is one CMCS example.

c) Tools must integrate with everyday scientific workflow providing 'end to end' capability and compatibility. Will adaptable workflow technologies contribute to scientific community adoption, support, operation, etc. of knowledge grid infrastructure?

d) What are the killer apps that attract scientists to adopt new systems?

3) Adaptability: Change is rampant, and will keep happening.

a) Even perfectly developed and implemented new infrastructure will be adopted only over time, in an iterative and evolutionary fashion. Thus, 'perfectly developed' means iteratively, adaptively, and with progressively more complex features. During adoption, requirements will likely change, and we must be prepared to invest later efforts in unexpected directions.

b) By its very nature, Science is changing. Data (and metadata) have become an important product of science, but data itself is rapidly changing – larger data sets, more complex associations in models, evolution with community evaluation, etc. New 'high throughput' tools and facilities (MP computers, light sources) amplify the rate of change.

While technology enabling the infrastructure is rapidly evolving, users are reluctant to invest in a system that is not expected to survive longer than the next software cycle. How can we abstract longer-lived aspects of enabling infrastructure while continuously upgrading technology and adding capabilities? Will adoption of open standards and open source technologies provide the needed longevity of this new knowledge infrastructure?